DOCUMENT RESUME

ED 317 602 TM 014 700

AUTHOR Lunz, Mary E.; Stahl, John A.
TITLE Severity of Grading across Time Periods.
PUB DATE Apr 90
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Clinical Diagnosis; Comparative Testing; Difficulty Level; Essay Tests; *Evaluators; Goodness of Fit; *Grading; Higher Education; *Interrater Reliability; *Medical Students; Test Format
IDENTIFIERS Oral Examinations; Rasch Model

ABSTRACT

Three examinations administered to medical students were analyzed to determine differences among severities of judges' assessments and among grading periods. The examinations included essay, clinical, and oral forms of the tests. Twelve judges graded the three essays for 32 examinees during a 4-day grading session, which was divided into eight half-day grading periods. Eighteen judges graded the performance of 217 examinees on the clinical examination during a 2-day grading session that was divided into four grading periods. Forty-six judges graded the performance of 270 examinees on the oral examination during a day and a half grading session that was divided into three grading periods. An extension of the Rasch model was used to analyze facets for examinees, items, judges, and grading periods. This study focused only on judge severities and differences among grading periods, however. The system of links necessary to calibrate judge severities and grading periods as separate facets was adequate because judges had 16 primary protocols and some examinees in common. Data from each of the three examinations were analyzed using FACETS, a computer program for Rasch analysis of examinations with more than two facets. The FACETS program estimates objective and conjointly additive calibrations, standard errors, and fit statistics for each element of each facet in the examination. Significant variation in judge severities and some variation across grading periods were found on all three examinations. However, the fit statistics confirm that most judges are reasonably consistent in the application of their individual level of severity. Four data tables and four graphs are included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED317602

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

MARY E. LUNZ

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

Severity of Grading Across Time Periods

Mary E. Lunz

John A. Stahl

American Society of Clinical Pathologists

Paper presented at the annual meeting of the American Educational Research Association, Boston, MA., April, 1990.

TM014700

BEST COPY AVAILABLE

Severity of Grading Across Time Periods

Abstract

Three examinations which require judges to assess examinee performances were analyzed to determine differences among judge severities and grading periods. An extension of the Rasch model analyzed facets for examinees, items, judges and grading periods. Significant variation in judge severities and some variations across grading periods were found on all three examinations.

Severity of Grading Across Time Periods

Assessment of essay, oral, clinical or other examinee performances usually requires the intervention of a judge. The expectation is that examinee scores will be independent of the particular judges that grade the performance and the grading period. The reality, however, is usually more as Thurstone (1927) observed, that the discriminal process corresponding to a given stimulus varies among individuals.

The validity and reliability of examinations which require judges have been questioned because of judge subjectivity and potential bias (Hurley, 1982) related to judges. Attempts to improve uniformity among judges have included constructing structured items such as essays or oral protocols, standardizing grading criteria and administration procedures, and providing extensive judge training. But these efforts have served only to direct the attention of judges, not to control the subjectivity of their assessments.

Inconsistency among judges has been studied extensively. Littlefield, et al (1981) compared the ratings of various types of judges (i.e. faculty and residents) and found significant differences in their assessments of similar clerkships. A multiple choice examination was found to be more reliable than the clinical ratings. Lunz and Stahl (1990) found inter and intra judge inconsistency when pass/fail decisions about the same examinee performance were made using different scoring criteria. Cason and Cason (1984) postulated that the ratings received by a subject are a function of the subject's true ability and the rater's characteristics including the rater's resolving power, sensitivity and stringency. A significant rater stringency effect and a

significant student ability effect were found. de Gruijter (1984) demonstrated differences among judges using linear and nonlinear analysis models. Lunz, et al (1989 and 1990) found that judges demonstrate discernable levels of severity which effect examinee scores on oral and clinical examinations. The results of these studies support the premise that judges have unique standards which interact with the examination materials and examinee performances resulting in differing levels of severity.[1] Standardized grading criteria and administration procedures can define the examination process, but can not remove differences in judge severity.

Examinations which require judges are often graded during defined grading sessions with delimited grading periods. Assuming that examinee ability is randomly distributed across grading periods, and that examinee performances are randomly allocated among judges, it is possible that some examinee performances are more or less severely graded during some grading periods. Thus the time of grading within the grading session in addition to the overall severity of the judge may influence the grade awarded. Braun (1988) found a sizeable shift in the average score of essay readers from day 1 to day 2.

Differences in judge severity and differences among grading periods for three examinations, an essay examination, an oral examination and a clinical examination will be explored. These three examinations have several attributes in common. Judges are needed to assess examinee performances and the grading sessions have defined grading periods. The judges come to a specific location to do the grading. The grading periods are contiguous.

1 Severity is the term used to describe the unique perception of a judge in regard to the examination materials, the standards for competence and the application of the rating criteria.

The elapsed time between grading periods ranges between one hour and 12 hours (overnight).

Data

There is no overlap among the three examinations in regard to items, examinees or judges. If similar patterns of differences among judge severities and across grading periods are found for the three examinations this would imply that the patterns are not unique to a particular examination.

The essay examination required the examinee to write three essays (items) so that their skill in english composition could be evaluated. Twelve judges graded the three essays for all 32 examinees during a four day grading session which was divided into eight half-day grading periods. Essays were graded on a nine point scale with 9 as excellent and 0 as unacceptable. A total of 27 points represented a perfect score on all 3 essays. These data have complete overlap, that is, all judges graded all essays for all examinees sometime during the eight grading periods. There is no missing data.

The clinical examination required examinees to prepare 15 histology slides (items) to detailed specifications. Eighteen judges graded performances from 217 examinees during a 2 day grading session, divided into four grading periods. It was impossible for all judges to grade all slides for all examinees (15 x 217 = 3255 slides), so examinee performances were allocated to judges. This introduced the opportunity for the severity of the judge, as well as the grading period, to influence the grade assigned. A rotation system enabled each judge to grade each of the 15 slides sometime during the two day grading session and some combination of three judges to grade subsets of an examination. This created the system of links necessary to calibrate judge severities and grading periods as separate facets. Even

though there is missing data, all judges have all slides and some examinee performances in common.

A perfect score was 75 points (15 slides x 5 points = 75 points). There were three assessments for each slide. Quality of tissue cutting and processing were graded acceptable (1 point) or unacceptable (0 points). Tissue staining was graded on a four point grading scale: unacceptable (0), below average (1), average (2) and above average (3). This design did not have complete overlap of judges, items, or examinees but did have a series of links based on common examinee performances and common items which enabled a complete calibration of all elements on one common scale (for a more complete explanation of linking see Lunz, Wright and Linarce, 1990).

The oral examination required examinees to complete two twenty-minute interviews, each with a different judge. These interviews were face-to-face interactions between the examinee and the judge. Forty-six judges graded 270 examinees during a day and a half grading session, divided into three grading periods. Twenty-seven structured protocols (items) were used for the examination (16 primary and 11 make-up). Each protocol described the nature of a case. The examinee then acquired additional information from the judge until a diagnosis could be made or a treatment determined. A four point grading scale was used on which 0 was unacceptable, 1 was below average, 2 was satisfactory and 3 was excellent. A perfect score was 18 points (6 protocols x 3 points = 18 points).

It was impossible for all judges to grade all examinees (2 interviews x 270 examinees = 540 interviews), so examinees were allocated to judges. This introduced the opportunity for the severity of the judge and the grading period to influence the score assigned. A rotation system in which examinees

were interviewed by two different judges using different subsets of protocols enabled each judge to grade each of the 16 primary protocols during the first two grading periods.

The third grading period was reserved for "make-up" examinations. Examinees who were not determined to be clear passes or fails after two interviews were examined a third time with a different judge and different protocols during this third grading session. Eleven different protocols were used during this session by a subset of the judges. The judges knew these were "make-up" examinees.

The system of links necessary to calibrate judge severities and grading periods as separate facets was adequate because judges have the 16 primary protocols and some examinees in common. The overlap among judges, examinees and protocols is least definitively defined for this examination and there is missing data.

Methods

It is usually assumed that the results of an examination generalize so that sensible action can ensue. An examinee who passes an examination is certified as having demonstrated an acceptable level of skill and knowledge, regardless of the specific sample of essays, clinical slides or protocols and regardless of the particular judge or grading period.

A measurement model designed to analyze an examination with multiple facets must provide an analysis of each of the elements in each facet of the examination. The particular elements within each facet must be calibrated in a way that is independent of the local distributions of the elements in the other facets. Thus, the positioning of examinee measures must function as though independent of which judges, items or grading periods were encountered.

The two facet (dichotomous) Rasch model $\log\{P_{ni}/(1-P_{ni})\} = B_n - D_i$ (Rasch, 1960/1980) analyzes the two facets of item difficulty and examinee ability. An examination with three or more facets, will include facets for examinee ability and item difficulty as well as any other facets, such as judge severity or grading period, which may effect examinee scores.

An extension of the Rasch model to include all facets which are pertinent to an examination was developed by Linacre (1989). The probability of person n with ability B_n achieving rating step x on item i with difficulty D_i from judge j with severity C_j during grading period T_t is modeled as $\log(P_{nijtx}/P_{nijtx-1}) = (B_n - D_i - C_j - T_t - F_x)$ (See Appendix 1 for explanation). This extended Rasch model constructs a variable, measured in log-odds units (logits), that quantifies the elements within each facet so that quantitative comparisons among and within the facets are possible. Each facet is calibrated from the relevant observed scores and all but the examinee measure facet are centered at a common origin.

The positioning of elements within each facet provides the frame of reference for verifying the intended examination definition. Examinee measures (B_n) are ordered from highest to lowest, judge severities (C_j) are ordered from most to least severe, and any differences among grading periods (T_t) are observable. It is also possible to observe how the grading categories (F_x) are used by the judges and the ordering of the examination item difficulties. This study, however, focuses on differences in judge severities and differences among grading periods. The other facets are calibrated as part of the analysis, but will not be discussed.

Data from each of the three examinations were analyzed using FACETS (Linacre, 1988) a computer program for Rasch analysis of examinations with more than two facets. The FACETS program estimates objective and conjointly additive (Luce & Tukey, 1964) calibrations, standard errors and fit statistics for each element of each facet in the examination. The examinee raw scores are linearized and corrected for variations in the measured severities of the particular judges and grading periods encountered by an examinee. The importance of this correction depends on the overlap among judges, items and examinee performances. The more variable the combinations, the more important the correction to obtain objectivity.

The fit statistics evaluate the suitability of the data for the construction of a variable and identify inconsistency for any element of any facet. Consistency verifies that these data are appropriate for making measures (Wright & Stone, 1979 chapter 4 and Wright & Masters, 1982 chapter 5). The fit statistics for judges indicate the degree to which each judge is internally self-consistent (intra-judge consistency). Deviant judges can be flagged. Unexpected scores can be identified and their effect on examinee measures analyzed. The fit statistics for each grading period indicate the inter-judge consistency among judges during that grading period.

Two kinds of fit to the expectations of the model are reported. The infit statistic is an information weighted mean-square residual which is sensitive to an accumulation of central or inlying deviations. The outfit statistic is an unweighted mean-square residual which is sensitive to occasional outlying deviations. The expected value for the mean squares is one (1.0) and their asymptotic standard errors are approximately the square root of (2/d.f.) where d.f. is the number of independent replications on which

the corresponding estimate is based. The region of acceptable fit will be mean squares greater than 0.5 and less than 1.5. Judges, or grading periods with infits or outfits beyond these criteria will be flagged and reviewed carefully for unexpected deviations.

The elements in each facet are summarized by their estimated mean, standard deviation, reliability of element separation and corresponding chi-square for homogeneity. In most test situations, variation in examinee performance is expected. When all examinees take all items and all judges grade all examinees, the variations in judge severities do not produce unfair scores. But when judges are allocated to examinee performances and grading periods vary, variation in judge severities and grading periods can effect raw scores and should be accounted for before examinee measures are calculated.

Separation reliability (similar to the KR-20) is the proportion of the observed variance in item difficulties, examinee measures, judge severities and grading period estimates not due to measurement error (Wright & Masters, 1982 pp 91-94). The chi-square for homogeneity tests whether the judges can be regarded as sharing the same severity after allowing for measurement error. A significant chi-square indicates that the variation in judge severities exceeds the error of measurement.

To determine the effect of grading period on examinee measures, each examination was analyzed twice, first with grading period modelled as a facet and again without grading period modelled as a facet. The second analysis assumes that all grading periods are comparable. In the essay examination where there is complete overlap, grading period should have no effect on examinee measures. For the clinical and oral examinations, in which judges are allocated to examinees, grading period may have an observable effect.

Results

Tables 1, 2 and 3 show the judge severity calibrations in order of severity for the essay, clinical and oral examinations and the summary statistics. For all three examinations calibrated judge severities show a range well beyond that expected due to error of measurement. The range of judge severities for the essay exam is .45 to -.30 logits and separation reliability is .82. For the clinical examination the range of judge severities is 1.21 to -.97 logits and separation reliability is .95. For the oral examination the range of judge severities is 1.67 to -1.58 logits and separation reliability is .86. The chi-square analyses for all three examinations confirmed that judge severities were significantly different ($p<.00$).

The fit statistics show intra-judge consistency for most judges within their level of severity. On the essay exam (Table 1) judge 4 is more consistent than expected (.5 infit & outfit). Review of the data found that this judge limited his use of the rating scale to points 5, 6 and 7 of the nine points possible. Judge 12 verged on misfit (infit 1.4 and outfit 1.4). Judge 12 awarded some ratings that were unexpectedly low (1 or 2 points) given his overall grading pattern. The examinees who received the low scores from this judge received relatively low scores from the other judges as well. All judges graded all examinees on all essays, yet measurably different judge severities are observable.

On the clinical examination (Table 2) no judge was sufficiently inconsistent to be outside the region of acceptable fit. Judges, however, manifest measurably different levels of severity.

On the oral examination (Table 3), five judges, 10, 15, 19, 25, 36, show misfit. Review of their data revealed that these judges gave unexpectedly low grades to some examinees. Judges 10 and 19 graded make-up examinees during the third grading period. These judges gave lower than expected scores (given their grading patterns) to these less able examinees, which show as intra-judge inconsistencies. Judges 15, 25 and 36 each graded one examinee lower than expected on one protocol which caused the misfit. Again, judges manifest measurably different levels of severity.

These analyses show that judges, regardless of the examination, can vary significantly in their severities but are generally consistent in their application of their level of severity across examinees.

Table 4 shows the calibrations of the grading periods for the three examinations. For the clinical examination the judges are more severe in the second grading period but less severe in the fourth period. For the oral examination the judges are consistent across the three grading periods. For the essay examination the judges are more severe during the third grading period. In all three examinations, the judges became less severe toward the end of the session. The infit and outfit statistics show that the grading period data fit the model. Both infit and outfit are within the acceptable region indicating inter-judge consistency within each grading period.

A Chi-square analysis for homogeneity across grading periods found significant differences for the clinical examination ($x^2 = 61.33$, df = 3, $p<.00$) and the essay examination ($x^2 = 17.90$, df = 7 $p<.00$) showing that the severity of grading can change significantly across grading periods. There was not a significant difference across grading periods for the oral examination.

INSERT GRAPHS 1, 2, 3 3A ABOUT HERE

The examinee measures for the essay examination with grading period (time) calibrated and grading period (time) uncalibrated are presented in Graph 1. There is a near perfect relationship between the measures earned with and without time as a calibrated facet. This is because the examination had complete overlap of judges, items and examinees. All examinees were graded during all time periods removing any advantage due to time.

The examinee measures for the clinical examination with time calibrated and time uncalibrated are plotted in Graph 2. Two distinct groups of examinees can be observed, those who are penalized and those who have an advantage due to the grading period. The examinee measures on line A were penalized due to the grading period, while those on line B had an advantage.

The examinee measures for the oral examination with time calibrated and uncalibrated are plotted in Graph 3. There is a slight advantage for some examinees due to grading period, although the effect is not marked because there is no significant difference among grading periods. Graph 3A, an enlargement of the measures around .00, shows that some examinees may have been penalized slightly due to grading period. The effect is not large, but it could have an impact on a few pass/fail decisions.

Discussion

These data demonstrate that judges differ in their severities regardless of the examination. The fit statistics for all three examinations, however, confirm that most judges are reasonably consistent in the application of their

individual level of severity. When it is possible for all judges to grade all examinee performances across all grading periods, the unique effects of judge and grading period are neutralized, as in the essay examination. When, however, for reasons of money or time, it is necessary to allocate examinee performances to subsets of judges, the effects of judge severity and grading period become important. Correction for grading period and judge severity improves the examinee measures because it frees them from the effects of the particular judge and grading period encountered and makes them more objective.

The Rasch fit statistics flag deviant grading patterns so that they can be reviewed. Misfit focuses diagnostic study of the data and provides specific information which can be shared with the judge. Detailed information about inconsistent grading can stimulate judges to think about their grading patterns and may lead to improved consistency.

Short term effects such as fatigue and attitude may account for the changes across grading periods. One can imagine that at the beginning of a grading session, judges get "warmed up". After they get "warmed up" they grade seriously, perhaps more severely, for a while. But then, as the end of the session draws near they "ease up" a little. This is perhaps normal human behavior, but it may also penalize a subgroup of examinees.

Training judges and developing detailed definitions of the scoring system and criteria help standardize the examination. After all reasonable efforts have been made to train judges, differences in severities are still observable. A training session of 2 to 3 hours may not be able to change ingrained personal expectations. It may be more reasonable to compensate for differences among judges than to attempt to make them comparable.

The use of the Rasch model places responsibility on the analyst. There may be a danger that judge severities can be over or under calibrated thus making an unfair adjustment to an examinee measure. The misfit statistics

flag this possibility so the data can be reviewed. There is also the need to create a sound linking network of items, judges and examinees. The FACETS program calculates the error of measurement for each element calibration and examinee measure. This quantifies the possible error associated with the use of the calibration or measure for decision purposes.

Any subgroup of judges is unique, so examinees who happen to get more lenient judges have a raw score advantage over examinees who happen to get more severe judges. This inequity is well documented but has been ignored because reasonable tools for dealing with the problem were not available. The use of the extended Rasch model provides these missing tools. The whole process of dealing with examinations that require judges becomes less mystical, more quantitative and more understandable to both judges and psychometric experts.

TABLE 1

Severity of Judges on Essay Exam in
Order of Severity

	Judge Number	Score	Count of Essays	Logit Judge Severity	Error	Infit MnSq	Outfit MnSq
Most	1	296	96	0.45	0.08	0.8	0.8
Severe	3	337	96	0.18	0.08	1.0	1.0
	6	338	96	0.17	0.08	1.0	1.0
	5	348	96	0.10	0.08	0.7	0.7
	10	365	96	-0.01	0.08	0.7	0.7
	11	370	96	-0.03	0.08	1.2	1.1
	12	374	96	-0.06	0.08	1.4	1.4
	2	377	96	-0.08	0.08	1.1	1.1
	7	383	96	-0.12	0.08	1.1	1.1
	9	388	96	-0.15	0.08	1.0	1.0
Least	4	389	96	-0.15	0.08	0.5	0.5
Severe	8	412	96	-0.30	0.08	1.3	1.2
	Mean:	364.8	96.0	-0.00	0.08	1.0	1.0
	S.D.:	29.5	0.0	0.19	0.00	0.2	0.2

Fixed (all same) chi-square: 65.88 d.f.: 11 significance: 0.00
RMSE - root mean square error of judge calibrations = .08
Adj S.D. - square root of observed variance minus mean square error variance =.17
Separation = (Adj S.D.)/RMSE = 2.16
Separation reliability = $(\text{Separation})^2/1+(\text{Separation})^2$=.82

TABLE 2

Severity of Judges on Clinical Examination in Order of Severity

	Nu	Score	Count of Slides	Calib. Judge Severity	Model Error	Infit MnSq	Outfit MnSq
Most	10	78	75	1.21	0.19	0.8	0.8
Severe	14	102	90	1.08	0.18	0.7	0.7
	8	683	615	0.70	0.07	0.9	0.9
	1	157	135	0.38	0.15	1.0	1.2
	15	884	705	0.25	0.07	1.2	1.1
	13	1054	840	0.16	0.07	0.8	0.9
	3	276	210	0.14	0.14	1.2	1.1
	9	779	615	0.14	0.08	1.0	1.0
	16	976	750	0.02	0.07	1.0	0.9
	6	1003	750	-0.13	0.08	1.1	0.9
	7	1396	1035	-0.24	0.07	1.0	0.9
	2	285	210	-0.39	0.15	1.1	1.0
	4	985	705	-0.41	0.09	1.0	0.8
	11	1333	960	-0.41	0.07	1.2	1.0
	18	1078	780	-0.41	0.08	1.2	1.2
	12	886	630	-0.54	0.09	1.0	1.2
Least	5	814	570	-0.56	0.10	1.1	1.0
Severe	17	127	90	-0.97	0.24	1.2	0.9
Count:	18 Mean:	716.4	542.5	0.00	0.11	1.0	1.0
	S.D.:	422.0	310.5	0.56	0.05	0.2	0.1

RMSE 0.12 Adj S.D. 0.55 Separation 4.49 Reliability 0.95
Fixed (all same) chi-square: 382.04 d.f.: 17 significance: 0.00
(see Table 1 for definitions)

TABLE 3
Severity of Judges on Oral Examination in
Order of Severity

	Judge	Score	Count of Protocols	Judge Severity	Error	Infit MnSq	Outfit MnSq
Most	33	56	39	1.67	0.25	1.3	1.3
Severe	29	68	39	1.51	0.26	1.1	1.0
	23	87	39	1.40	0.32	0.6	0.5
	49	66	33	1.34	0.32	0.7	0.6
	39	75	36	1.32	0.30	1.4	1.4
	10	78	42	1.13	0.27	**2.0**	**2.2**
	26	75	36	1.08	0.30	0.9	0.9
	40	67	33	1.04	0.31	0.8	0.9
	7	76	39	0.74	0.27	0.6	0.6
	43	38	24	0.69	0.37	0.7	1.2
	12	79	33	0.65	0.34	0.9	0.8
	18	75	36	0.64	0.30	0.6	0.6
	46	74	36	0.63	0.29	0.9	0.8
	8	81	42	0.52	0.26	0.8	0.7
	31	77	42	0.48	0.26	0.6	0.5
	15	89	39	0.46	0.30	**1.6**	**1.6**
	16	62	33	0.45	0.30	0.6	0.6
	47	88	42	0.39	0.28	1.0	1.0
	3	66	33	0.36	0.31	1.2	1.0
	44	82	40	0.22	0.28	1.0	0.9
	34	86	45	-0.02	0.25	1.0	1.0
	30	77	36	-0.14	0.31	1.1	1.1
	45	99	45	-0.20	0.27	1.0	0.8
	48	81	39	-0.23	0.29	1.0	0.9
	35	68	33	-0.25	0.31	1.0	1.1
	37	75	36	-0.28	0.31	1.0	1.0
	50	78	33	-0.32	0.34	1.1	1.2
	42	82	36	-0.34	0.33	0.8	0.7
	11	72	37	-0.42	0.29	0.7	0.8
	19	88	42	-0.46	0.27	**1.5**	**1.6**
	32	79	33	-0.51	0.34	0.8	0.8
	41	66	33	-0.54	0.31	1.2	1.1
	2	73	30	-0.55	0.37	1.2	1.3
	1	102	48	-0.68	0.26	0.7	0.7
	14	68	33	-0.68	0.31	1.1	1.1
	51	78	36	-0.75	0.31	0.6	0.6
	13	96	42	-0.77	0.29	1.0	0.9
	6	77	36	-0.83	0.31	0.7	0.8
	20	68	30	-0.83	0.35	1.2	1.1
	28	73	33	-0.85	0.33	0.9	0.9
	36	68	30	-0.91	0.36	**1.4**	**1.6**
	52	70	30	-1.10	0.36	0.7	0.7
	21	79	36	-1.11	0.30	0.7	0.7
	53	63	27	-1.17	0.38	0.7	0.6
Least	25	81	36	-1.20	0.33	**1.6**	**1.7**
Severe	17	91	36	-1.58	0.38	1.3	1.1
	Mean:	76.0	36.2	0.00	0.31	1.0	1.0
	S.D.:	11.1	4.8	0.83	0.03	0.3	0.3

RMSE 0.31 Adj S.D. 0.78 Separation 2.51 Reliability 0.86
Fixed (all same) chi-square: 345.12 d.f.: 45 significance:0.00
(see Table 1 for definitions)

TABLE 4

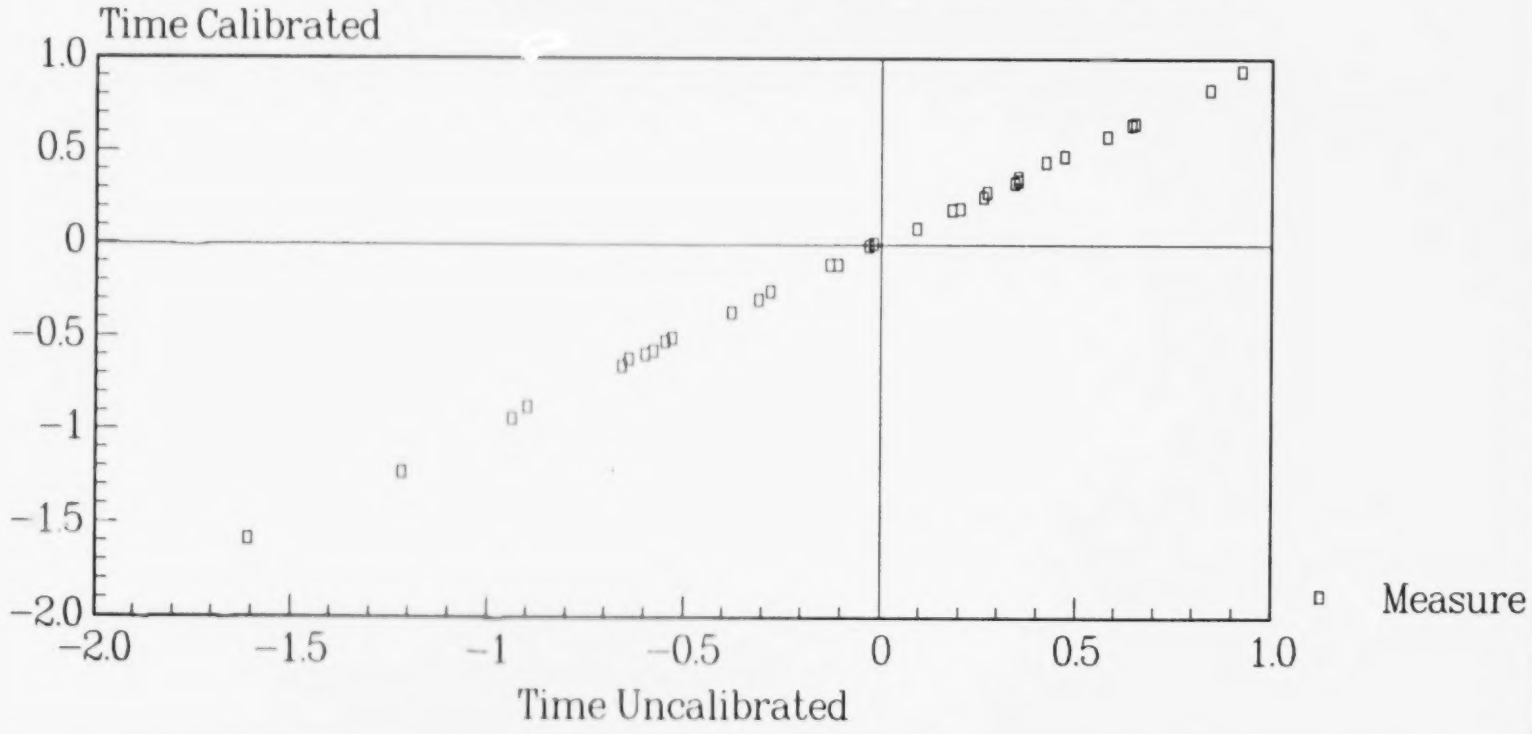
Grading Severity Calibrations Across Time Periods
For Clinical, Oral, Essay Examinations

Examination	Time Period	Grading Severity Calibrations*	SE	Consistency Infit MnSq.	Consistency Outfit MnSq.
Clinical	1 morning	-.03	.04	1.1	1.1
	2 afternoon	.30**	.04	1.0	.9
	3 morning	-.05	.03	1.0	1.0
	4 afternoon	-.22**	.06	1.0	.9
Oral	1 morning	.06	.06	.9	.9
	2 afternoon	-.05	.07	1.1	1.1
	3 morning	-.01	.17	1.2	1.1
Essay	1 morning	.05	.07	1.1	1.1
	2 afternoon	.05	.07	1.1	1.1
	3 morning	.20**	.07	.9	.9
	4 afternoon	.02	.07	1.1	1.1
	5 morning	-.02	.07	1.1	1.1
	6 afternoon	-.11	.07	1.0	1.0
	7 morning	-.11	.07	.8	.8
	8 afternoon	-.08	.07	1.1	1.1

* Positive calibration = more severe grading;
negative calibration = more lenient grading

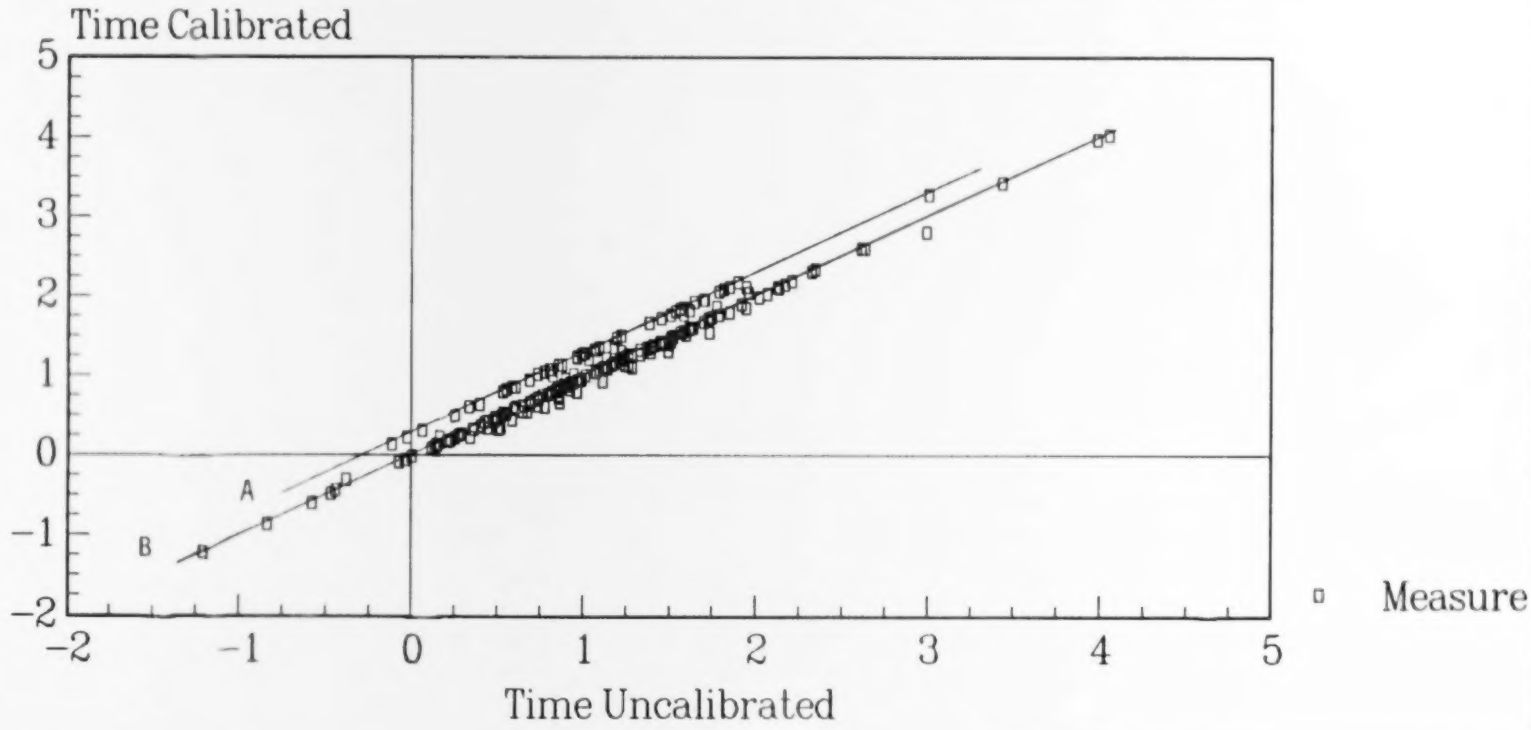
** Statistically significant difference, chi-square analysis

GRAPH 1
ESSAY EXAMINATION MEASURES
TIME CALIBRATED VS. TIME UNCALIBRATED



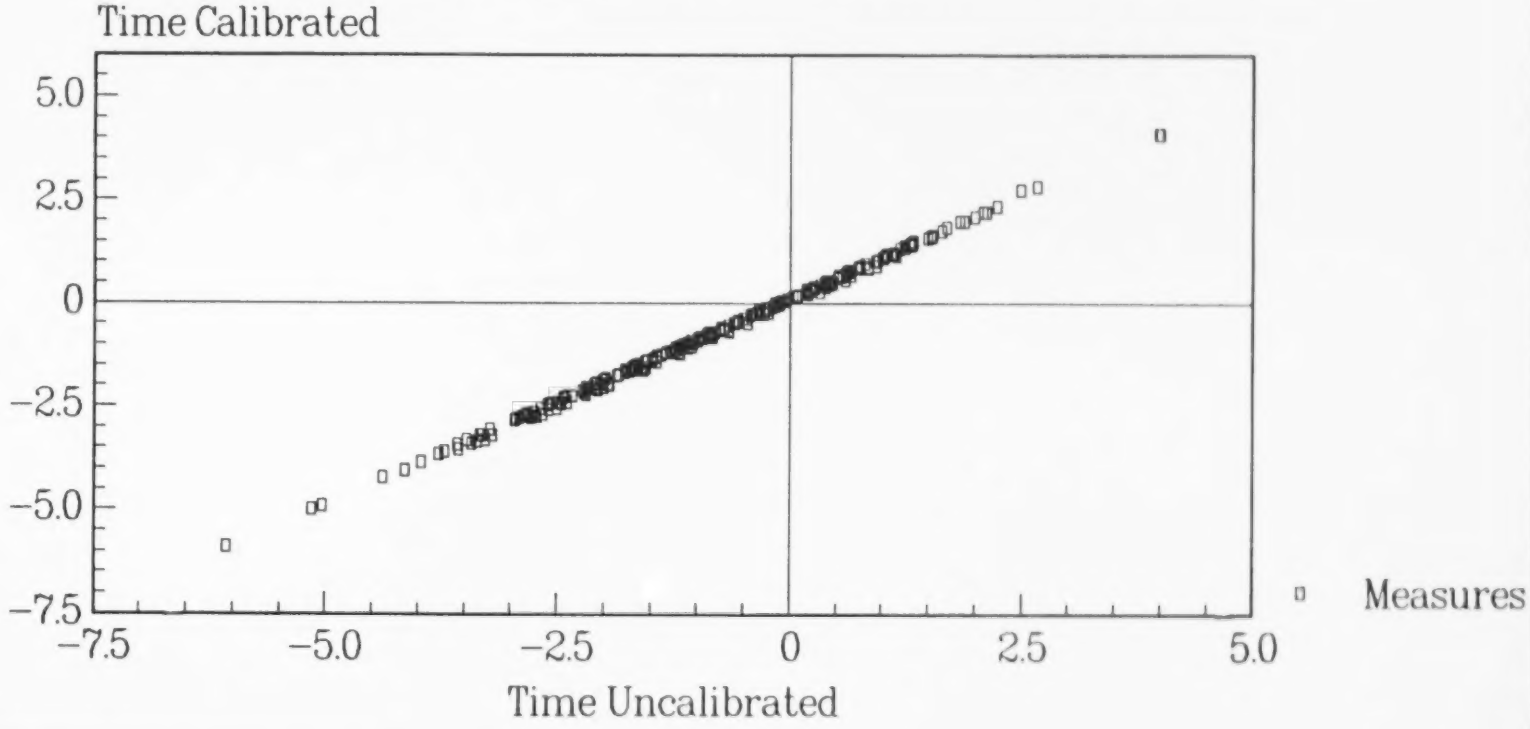
ESSAY

GRAPH 2
CLINICAL EXAM MEASURES
TIME CALIBRATED VS. TIME UNCALIBRATED



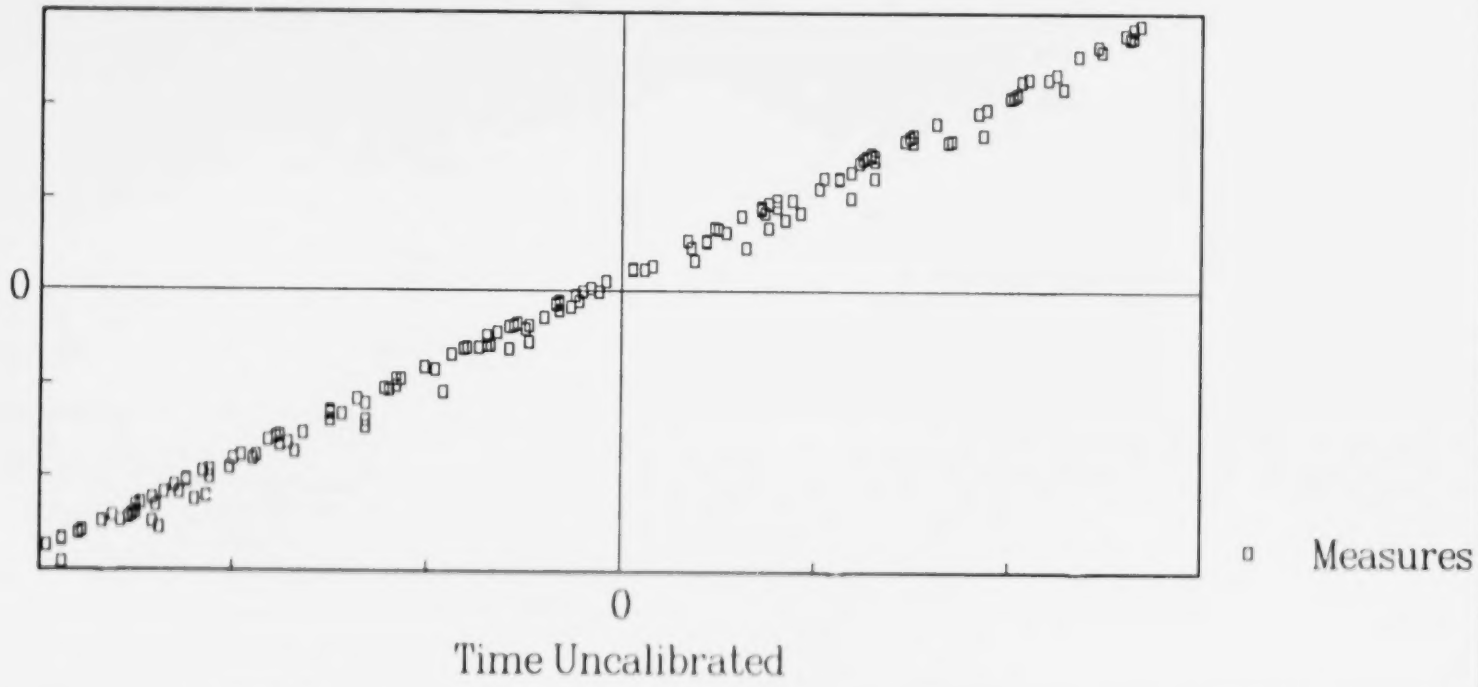
PRACTEST

GRAPH 3
ORAL EXAMINATION MEASURES
TIME CALIBRATED VS. TIME UNCALIBRATED



ORALT2

GRAPH 3A
ORAL EXAMINATION MEASURES
TIME CALIBRATED VS. TIME UNCALIBRATED



ORALXPAN

23

APPENDIX 1

$$\log\left[\frac{P_{nijtk}}{P_{nijtk-1}}\right] = B_n - D_i - C_j - T_t - F_k$$

P_{nijtk} - Probability of person n being given grade k by judge j on item i at time t

$P_{nijtk-1}$ - Probability of person n being given grade k-1 by judge j on item i at time t

B_n - ability of person n
D_i - difficulty of item i
C_j - severity of judge j
T_t - stringency of time t
F_{mk} - difficulty of grading step k relative to step k-1

The probability of a performance (B_n) earning a particular measure depends upon the rating (k) awarded and the additive effects of the difficulty of the item (D_i) the severity of the judge (C_j), the grading period (T_t) and the difficulty of the grading step (F_k). Misfit statistics identify the particular gradings which are improbable and provide a check on the technical validity of the measures. This study focuses on judge severity and grading period, however, the other facets are also included in the equation to produce more precise estimates.

References

Braun, H.I. (1988). Understanding scoring reliability: experiments in calibrating essay readers. *Journal of Educational Statistics*, 13, 1-18.

Cason, G.J. and Cason, C.L. (1984). A deterministic theory of clinical performance rating. *Evaluations and the Health Professions*, 7, 221-247.

de Gruijter, D.N.M. (1984). Two simple models for rater effects. *Applied Measurement in Education*, Ia Press.

Hurley, J.J. (1982). The Part II certifying examination in dermatology: An Assessment of Interpretive Skills in J.S. Lloyd (ed.). *Evaluation of Doncognitive Skills and Clinical Performance*. Chicago: American Board of Medical Specialties.

Linacre, J.M. (1989). *Multi-faceted Rasch Measurement*. Chicago: MESA Press.

Littlefield, J.H., Harrington, J.T., Anthracite, N.E. and Garman, R.E. (1981). A description and four-year analysis of clinical clerkship evaluation system. *Journal of Medical Education*, 56, 334-340.

Luce, .E. and Tukey, J.W. (1964). Simultaneous conjoint measurement, a new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1-27.

Lunz, M.E., Wright, B.D. and Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*. In Press.

Lunz, M.E., Stahl, J.A., Wright, B.D., Linacre, J.M. (1989). *Variations Among Examiners and Protocols on Oral Examinations*. Paper presented at the annual meeting of the American Educational Research Association. San Francisco, CA. (ERIC DOCUMENT, TM012988).

Lunz, M.E. and Stahl, J.A. (1990). A comparison of intra and interjudge decision consistency using analytic and holistic scoring criteria. *Journal of Allied Health*, In Press.

Rasch, G. (1960/1980). *Probablistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press.

Thurstone, L.L. (1927). A law of comparative judgment. *Psychological Review*, 34, 273-286.

Wright, B.D. and Masters, G.N. (1982). *Rating Scale Analysis* Chicago: MESA Press.

Wright, B.D. and Stone, M.H. (1979). *Best Test Design*. Chicago: MESA Press.